第46卷 第11期
2023年11月

干旱区地理
ARID LAND GEOGRAPHY

Vol. 46 No. 11
Nov. 2023

基于特征波段选择和机器学习的陆地棉叶片水分估算

崔锦涛[1]， 买买提·沙吾提[1,2,3]

（1. 新疆大学地理与遥感科学学院，新疆 乌鲁木齐 830046；2. 新疆绿洲生态重点实验室，新疆 乌鲁木齐 830046；3. 智慧城市与环境建模自治区普通高校重点实验室，新疆 乌鲁木齐 830046）

摘 要： 棉花叶片含水量的及时准确监测对于评价棉花生长状态具有重要作用。为了精准估算棉花叶片含水量，以新疆渭干河−库车河三角洲绿洲田间尺度上棉花叶片的高光谱数据和叶片水分数据为基础，采用分数阶微分对原始光谱进行处理，通过相关系数分析法、竞争性自适应重加权采样算法（Competitive adaptive reweighted sampling，CARS）、连续投影算法（Successive projections algorithm，SPA）、遗传算法（Genetic algorithm，GA）、蒙特卡罗无信息变量消除算法（Monte Carlo uninformative variables elimination，MC-UVE）以及将CARS与SPA耦合等方法筛选特征波段，采用基于鲸鱼优化算法（Whale optimization algorithm，WOA）改进随机森林回归（Random forest regression，RFR）建立全波段和特征波段的叶片水分含量反演模型，并使用独立样本进行验证分析。结果表明：（1）不同特征波段筛选方法得到的波段数量与位置不同，其中MC-UVE所得特征波段数量为8个，CARS所得特征波段数量为38个。SPA、GA与CARS-SPA方法中特征波段位置较为一致，基本集中在近红外的950~1050 nm范围内。（2）CARS-SPA-WOA-RFR模型反演效果最好，模型预测值决定系数（R^2）=0.93，均方根误差（Root mean square error，RMSE）=0.032。最终构建的模型可为准确快速地监测棉花旱情以及精准灌溉提供决策依据。

关 键 词： 光谱；叶片含水量；特征波段选择；机器学习

文章编号： 1000－6060(2023)11－1836－12(1836～1847)

我国作为最大的棉花生产国，主要生产区域集中在新疆棉区、黄河流域棉区、长江流域棉区[1]。其中，新疆棉花种植规模和产量均居全国首位[2]。棉花生长过程中通过叶片进行光合作用产生其所需要的能量，而叶片含水量对于监测生理状态、评估作物长势、反映土壤墒情等具有重要作用。因此，快速有效地获取叶片水分含量对于干旱半干旱区棉花生长、产量评估、旱情评价等具有重要意义。

高光谱遥感技术凭借快速准确和无损的优势克服了传统的实验室测量叶片含水量数据耗时耗力、具有破坏性、无法快速且大面积地获取棉田的水分数据的不足，已被广泛运用于作物水分反演中，并取得了许多成果[3-4]。以往研究大多采用竞争性自适应重加权采样算法（Competitive adaptive reweighted sampling，CARS）、连续投影算法（Successive projections algorithm，SPA）、遗传算法（Genetic algorithm，GA）、随机森林（Random forest，RF）与蒙特卡罗无信息变量消除算法（Monte Carlo uninformative variables elimination，MC-UVE）等方法筛选特征波段或构建植被指数，借助偏最小二乘回归（Partial least squares regression，PLSR）、支持向量机回归（Support vector machine regression，SVR）、反向传播（Back propagation，BP）和RF等机器学习方法建立反演模型[5]。如Sun等[6]采用SPA、CARS、逐步回归（Stepwise regression，SR）以及耦合方法筛选特征波长，使用多元线性回归（Multivariable linear regression，MLR）建立茶叶叶片水分含量的反演模型，结果表明CARS-SR方法所构建模型效果最优。Li等[7]

收稿日期： 2022-12-15； **修订日期：** 2023-01-17
基金项目： 新疆自然科学计划（自然科学基金）联合基金项目（2021D01C055）资助
作者简介： 崔锦涛（1997-），男，硕士研究生，主要从事干旱区资源环境及农业遥感应用方面的研究. E-mail: 107552101099@stu.xju.edu.cn
通讯作者： 买买提·沙吾提（1976-），男，博士，副教授，主要从事干旱区资源环境及农业遥感应用方面的研究. E-mail: korxat@xju.edu.cn

采用CARS、SPA、RF与联合间隔偏最小二乘(Synergy interval partial least squares，SiPLS)方法筛选特征波段，运用最小二乘支持向量机(Least-squares support-vector machines，LSSVM)模型反演柠檬叶片的叶绿素，决定系数(R^2)达0.94。杨宝华等[8]使用MC-UVE、随机蛙跳、CARS与移动窗口偏最小二乘(Moving window partial least squares，MWPLS)方法筛选特征波段，使用BP、SVR与径向基函数(Radial basis function，RBF)建立小麦冠层氮含量的估测模型，结果表明CARS-RBF模型R^2为0.998。以上研究表明，CARS在筛选特征波段方面能取得较好的效果。而张文旭等[9]和易翔等[10]借助SPA筛选棉花叶片氮素含量与地上部生物量的特征波段，利用PLSR建立反演模型取得了较高的反演精度。众多学者利用不同的筛选方法对作物属性的定量监测进行了大量的研究，并且取得了显著的成果。因此，选择合适的特征波段筛选方法对于利用光谱反演作物水分尤为重要。

相较于传统的MLR、SR以及岭回归等预测精度低、易受变量及样本数量影响的缺点，机器学习算法有效弥补了不足，比如SVM和RF模型广泛应用于叶片水分[3]、叶绿素[11]以及土壤有机质[12]和pH[13]相关研究中。与此同时各种优化算法也发展迅速，比如粒子群优化算法(Particle swarm optimization algorithm，PSO)、极限学习机(Extreme learning machine，ELM)和鲸鱼优化算法(Whale optimization algorithm，WOA)等广泛应用于传统机器学习的优化中，其中WOA具有原理简单易懂、需要调节参数少、精度高、收敛过程迅速和不易陷入局部最优等特点[14]。如Zhou等[15]使用WOA对SVM模型进行参数寻优，其改进后的WOA-SVM模型与SVM、ANN模型相比能够获得很好的建模精度。Zhao等[16]运用WOA进行参数优化，与LSSVM相结合实现了对$PM_{2.5}$的准确预测。

针对当前区域的棉花水分研究中大多采用线性回归、SVR与随机森林回归(Random forest regression，RFR)方法[17-18]，WOA在基于光谱测算作物水分方面的应用鲜有报道。尤其是反演精度和拟合效果等方面还未进行深入的研究，因此将WOA应用于作物含水量的研究尤为重要。本研究以棉花叶片水分含量为研究对象，使用分数阶微分对光谱进行预处理，采用6种特征波段筛选方法，基于WOA改进RFR算法(WOA-RFR)构建棉花叶片含水量的反演模型，并通过独立样本对模型进行验证，为快速准确地监测棉花叶片含水量提供技术支持。

1 材料与方法

1.1 研究区概况

渭干河-库车河三角洲绿洲(简称渭-库绿洲)，位于新疆阿克苏地区，地处塔里木盆地北部，天山南部，是一个典型的山前冲积扇平原(图1)。研究区属于典型的大陆性暖温带干旱气候，年均气温为10.5~11.4 ℃，山区多年平均降水量为243.0 mm，平原区多年平均降水量为46.5 mm，平原区年平均蒸发量为1374 mm，具有气候干旱、降水稀少、风沙频繁等特点[19]。2011年统计数据显示，渭-库绿洲棉花种植面积分别占新疆全疆、阿克苏地区的8.56%、38.2%，产量分别占8.41%、40.34%，是新疆主要的棉花生产区域之一[20]。本次试验田位于库车市中东部的乌尊镇，地理坐标介于41°31′29.65″~41°49′6.73″N，83°00′13.7″~83°19′15.95″E之间。研究区主要经济作物有棉花、辣椒等，种植品种为陆地棉，种植行距70 cm，株距10 cm；按照新疆农业农村厅农业技术推广总站的要求合理施肥和管理，全生育期一般喷施缩节胺5次左右，分别在苗期、二叶期、头水前以及打顶后5 d及12 d各喷施一次，6—8月每隔7 d、8 d滴灌一次，8月25日前结束灌水施肥。

1.2 叶片光谱测定及处理

实验室团队于2018年7月5—9日在新疆阿克苏地区库车市乌尊镇开展叶片光谱测定试验。使用ASD Field Spec Hand Held便携式光谱仪测定棉花叶片高光谱数据，波长范围为325~1075 nm，光谱测量时选择太阳辐射相对稳定且无风无云的天气，测量时间为11:00—14:00，将探头置于棉花叶片垂直上方25 cm处，设置光谱扫描时间为8 s，每个样点测6次，每3次光谱测定后利用白板进行标定，以消除光线变化对光谱的影响[11]。测定样本选择无病虫害、冠层生长均匀一致的棉花冠层，选择其中第二或者第三片棉叶进行光谱测定。测得反射率数据使用ViewSpec PRO计算出每个样点6条曲线的平均值作为该点的反射光谱，共采集到100个花铃期棉花叶片样本，命名为数据集Ⅰ。为了更好地验证所选模型的稳定性，选择2021年5月26日至6月

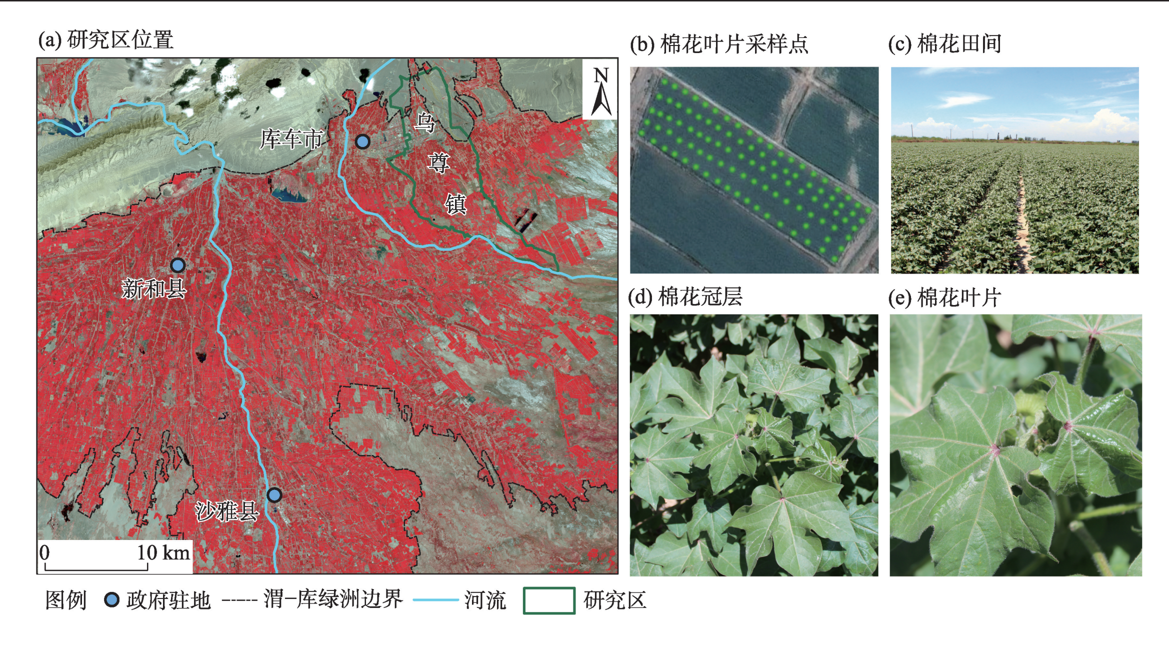


图1 研究区位置与采样点分布

Fig. 1 Study area location and sampling point distribution

3日测定的125个新陆早35号苗期棉花叶片样本作为数据集Ⅱ进行再次验证。

分数阶微分在图像增强处理和信号分析等领域被广泛的使用，多用来细化光谱信息。其原理是将整数阶微分的阶数以0.1为步长扩展至0~2阶。常用的分数阶微分包括Riemann-Liouville、Caputo和Grünwald-Letnikov 3种类型，其中Grünwald-Letnikov定义的微分形式较为常用，故用其对叶片高光谱数据进行处理[21-22]。微分公式为：

$$\frac{d^{\alpha} f(\lambda)}{d\lambda^{\alpha}} \approx f(\lambda)+(-\alpha)f(\lambda-1)+\frac{-\alpha(-\alpha+1)}{2}f(\lambda-2)+\cdots+\frac{\Gamma(-\alpha+1)}{m!\Gamma(-\alpha+1)}f(\lambda-m) \quad (1)$$

式中：Γ 为Gamma函数；α 为任意阶数；λ 为对应的波长点；$f(\lambda)$ 为 λ 的函数；d 为分数阶微分的下限；m 为微分上下限之差。 α 为小数时，则为分数阶微分变换。

1.3 叶片含水量测定

采集光谱测定后的叶片立刻使用0.001 g的电子天平称取每个叶片重量，记为鲜重。然后立刻置于保鲜袋内，当日野外试验完成后，在实验室内进行样本的干燥处理。使用烘箱在105 ℃下杀青30 min，然后以恒温80 ℃烘干，直至恒重后再测量叶片干重。叶片含水量（LWC）计算公式如下：

$$LWC = \frac{FW - DW}{FW} \times 100\% \quad (2)$$

式中：FW为棉花叶片鲜重（g）；DW为棉花叶片干重（g）。

1.4 建模集和验证集的划分

使用随机抽样方法将采集的样本进行分类，数据集Ⅰ中70个作为建模集，剩余30个作为验证集。数据集Ⅱ中按相同比例抽取38个为验证集。各分组统计量见表1。经统计分析，在数据集Ⅰ和Ⅱ中，样本集、建模集和验证集各项指标都比较接

表1 棉花叶片含水量统计分析

Tab. 1 Descriptive statistic of cotton leaf water content

数据类型	样本类型	样本数/个	叶片水分含量/%		
			均值	标准偏差	变异系数
数据集Ⅰ	建模集	70	78.34	0.043	5.54
	验证集	30	77.30	0.042	5.40
	样本集	100	78.03	0.043	5.50
数据集Ⅱ	验证集	38	79.00	0.074	9.35
	样本集	125	77.79	0.080	10.26

近，表明样本划分满足随机性和代表性，符合光谱技术建模的要求。

1.5 光谱特征变量筛选方法

为评价各种筛选方法对叶片水分反演的效果与作用，采用6种特征变量筛选方法以及3种机器学习模型进行反演（表2、图2）。为了充分发挥分数阶微分在细化光谱信息中的作用，以提高建模效果，本文采用分数阶微分处理原始光谱数据，以相关系数（CC）分析法筛选特征波段。而CARS、SPA、GA、MC-UVE以及CARS-SPA则使用原始数据进行筛选。

1.6 模型构建与验证

本研究所用的机器学习方法为WOA-RFR，并与目前常用的SVR和RFR模型对比分析。

WOA是Mirjalili等[27]于2016年提出，该算法仿照座头鲸的泡泡网觅食方法，通过收缩包围、螺旋位置更新以及随机捕食行为捕猎，进而建立数学模型，是一种群体智能算法。

1.6.1 包围猎物 在捕猎过程中，以距离目标猎物最近的座头鲸位置为最优位置，其他座头鲸向该位置运动以完成包围猎物。该表达式为：

$$D = \left| CX^*(t) - X(t) \right| \tag{3}$$

$$X(t+1) = X^*(t) - AD \tag{4}$$

式中：D为最优个体位置与当前个体位置的距离；$X^*(t)$为当前最优解的位置向量；$X(t)$为当前解的位置向量；$X(t+1)$为迭代位置向量；t为迭代次数；A和C为参数向量，其具体数学表达式分别为：

$$A = 2ar_1 - a \tag{5}$$

$$C = 2r_2 \tag{6}$$

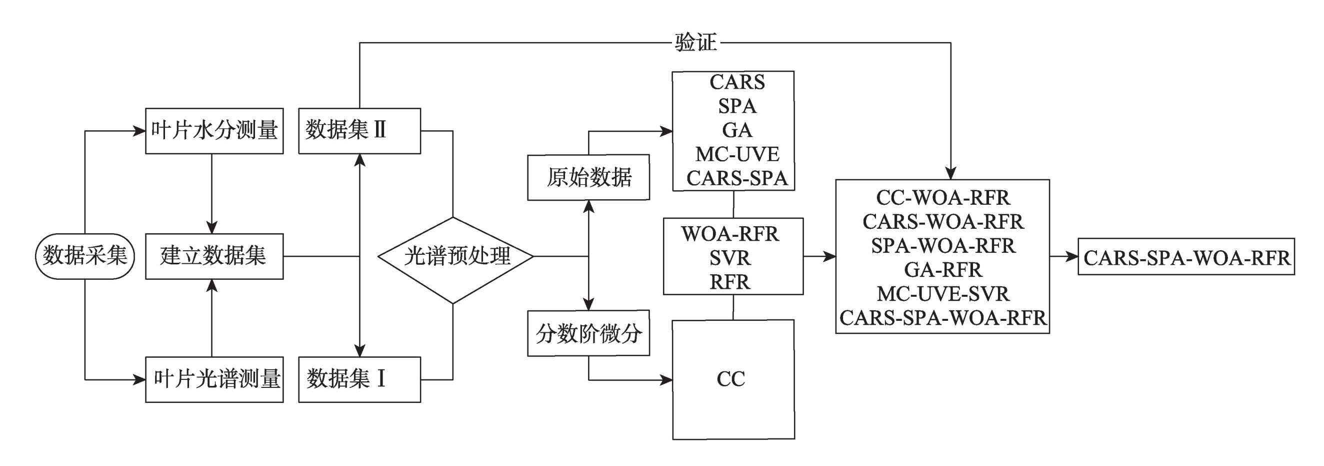
式中：r_1与r_2为值在[0, 1]中的随机向量；a为迭代搜索中从2线性减小到0，其数学表达式为：

表2 光谱特征变量筛选方法

Tab. 2 Spectral characteristic variable screening methods

变量筛选方法	描述
CC	运算效率高，过程简单。变量间存在共线性。
CARS	有效去除自相关性高的波段，适合高维数据的筛选[23]。变量间存在共线性，选择波段稳定性低。
SPA	变量间冗余少，共线性最小，缩短建模时间[24]。没有考虑所有特征波长之间的共线性[25]；挑选特征变量过程中倾向于选择共线性较小的变量点而不是有效变量点[25]。
GA	具有全局优化能力。但需要多次运算以确定最佳变量子集。
MC-UVE	稳定性较高。需要定义阈值，导致变量数目改变。
CARS-SPA	进一步剔除冗余信息，提取出有效波段且多重共线性较低，运算效率较高[6]。筛选变量较少，容易丢失关键信息。特征波长集建模效果受粗选算法结果的影响较大[26]。

注：CC为相关系数；CARS为竞争性自适应重加权采样算法；SPA为连续投影算法；GA为遗产算法；MC-UVE为蒙特卡罗无信息变量消除算法。下同。



注：CC为相关系数；CARS为竞争性自适应重加权采样算法；SPA为连续投影算法；GA为遗传算法；MC-UVE为蒙特卡罗无信息变量消除算法；WOA鲸鱼优化算法；SVR为支持向量机回归；RFR为随机森林回归。下同。

图2 实验及模型计算流程图

Fig. 2 Flow chart of calculation for experiences and models

$$a = 2 - 2\frac{t}{t_{\max}} \tag{7}$$

式中：$t_{\max}$ 为最大迭代次数。

1.6.2 螺旋泡泡网攻击 通过不断收缩包围机制和更新螺旋位置两种机制进行攻击，将座头鲸的泡沫网行为建立数字模型。首先计算与猎物之间的距离，建立螺旋运动的数学模型。表达式为：

$$\begin{cases} X(t+1) = X^*(t) + D_{\mathrm{P}}\mathrm{e}^{bl}\cos(2\pi l) \\ D_{\mathrm{P}} = \left| X^*(t) - X(t) \right| \end{cases} \tag{8}$$

式中：D_{P} 为猎物与鲸鱼之间的距离；b 为对数螺旋系数；l 为(−1, 1)区间内的随机数。

1.6.3 搜索猎物 座头鲸除了利用螺旋泡泡网搜索目标猎物外，还会根据与猎物之间的位置进行随机运动寻找猎物。该行为根据向量 A 的变化进行选择，A 在(−1,1)之外取随机值，使得鲸鱼去搜寻更合适的目标，从而提高全局寻优的能力。数学模型为：

$$D = \left| CX_{\mathrm{rand}}(t) - X(t) \right| \tag{9}$$

$$X(t+1) = X_{\mathrm{rand}}(t) - AD \tag{10}$$

式中：X_{rand} 为座头鲸群体中被随机选中的个体的位置向量[28]。

根据实验结果，在本研究中设置鲸鱼进化参数为30，种群规模 N=20。进而优化回归树棵数、回归树最大深度、节点划分所需最小样本数、叶子节点最少样本数。

模型预测精度选取 R^2、建模集和验证集均方根误差（Root mean square error，RMSE）作为衡量标准。R^2 越接近1，表明模型精度越高，拟合效果越好，预测值和实测值之间的相关性越强。RMSE是衡量模型预测值误差大小的指标，值越小模型的预算估测能力越好。计算公式如下：

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{11}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{12}$$

式中：$\hat{y}_i$ 为模型预测值；y_i 为实测值；$\bar{y}$ 为实测值的平均值；n 为样本个数。

2 结果与分析

2.1 光谱特征变量筛选方法结果与分析

本研究中利用CC、CARS、SPA、GA、MC-UVE和CARS-SPA方法对光谱数据进行特征波段选择。

CC：对原始光谱反射率以0.2为间隔，在0~2阶进行微分处理。由图3可知，光谱数据经分数阶微分变换后，在8个阶次中有通过0.01显著性水平检验的波段，CC绝对值最大为0.379。叶片原始光谱和分数阶微分处理后光谱与叶片含水量逐波段做Pearson相关分析（以1.3阶为例），得出CC在各波长上的分布图（图4）。原始光谱各波段均未能通过0.01水平的显著性检验，因此，采用分数阶微分方法对原始光谱进行处理，以提高相关性水平。经过分数阶微分处理，避免了信息遗漏，对消除噪声有一定的效果，提升了光谱的表达能力。

CARS：随着不断迭代，特征波长数量减小速度逐渐变缓，表明CARS算法在筛选特征波段中具有

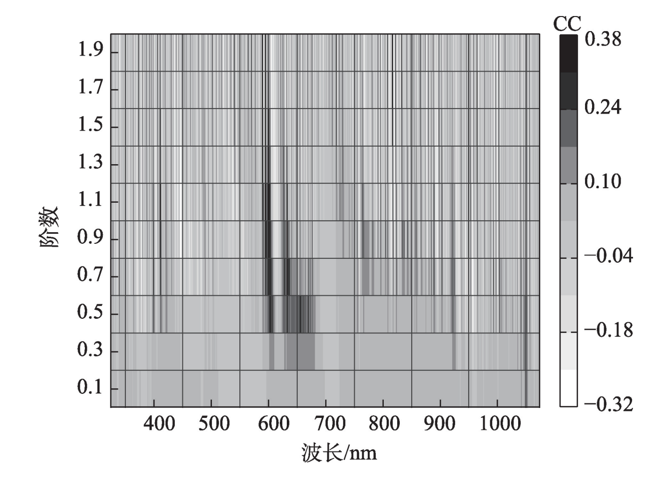


图3 分数阶微分转换光谱与棉花水分含量之间的相关性

Fig. 3 Pearson correlations between cotton leaf water content and fractional-order derivative spectra

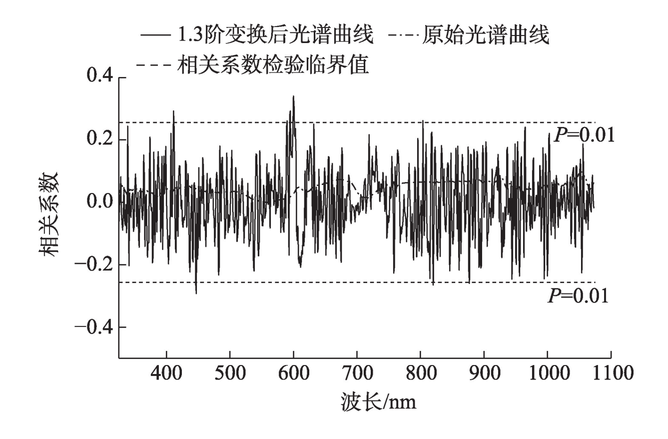


图4 不同光谱变换形式的相关性分析

Fig. 4 Correlation analysis of different spectral transformation forms

“粗选”和“精选”2个阶段(图5a)。当采样次数约为51次时十折交叉验证均方根误差(RMSECV)最小，即图中竖线所对应值(图5b)。回归系数值不断变化，表明运算过程中先剔除与叶片含水量相关性较弱的波段，而后又剔除了与叶片含水量相关性强的波段(图5c)。分析发现，当采样次数为第51次时，RMSECV最小，共有38个波段，提取的波段数量仅占原始波段数的5%，有效降低了光谱信息的冗余。

SPA：图6a中方格所示为最优子集中包含的样本数，图6b中方格所示为最优子集的波段位置。随着筛选变量数量的增加，RMSE迅速上升，当变量数为10时，RMSE趋于稳定，为0.032，表明其为最优子集。通过SPA算法共提取10个特征波长，占原始波段的1.3%。

GA：基于GA算法的特征光谱筛选结果(图7)，本研究采用50次运行GA算法，选取结果中出现频率较高的10个波长，作为最终的特征波段子集。

MC-UVE：基于MC-UVE算法的波长变量筛选结果(图8)，通过对全波段逐个计算稳定性值，最终选出8个波长作为特征波段子集。

CARS-SPA：由图5可知，经过CARS算法计算后，特征波长变量较多，波长之间有存在共线性的可能，因此耦合CARS-SPA模型。图9a中方格所示为最优子集中包含的样本数，图9b中方格所示为最优子集的波段位置。当波段数为10时，RMSE趋于稳定，为0.144，表明其为最优子集。通过CARS-SPA运算以后，共提取10个特征波长，占原始波段的1.3%，与CARS筛选变量相比，进一步减少了计算量。

上述6种变量筛选方法所得波段的位置分布如图10所示。通过分析发现，MC-UVE筛选变量最少共8个，CARS最多共38个。SPA、GA、CARS-SPA所筛选波段较为集中，主要分布在紫外区和近红外区，其近红外区为叶片含水量的敏感区域。

2.2 模型建立与分析

为研究不同变量筛选方法对模型精度的影响，对全波段以及6种筛选方法的结果构建WOA-RFR

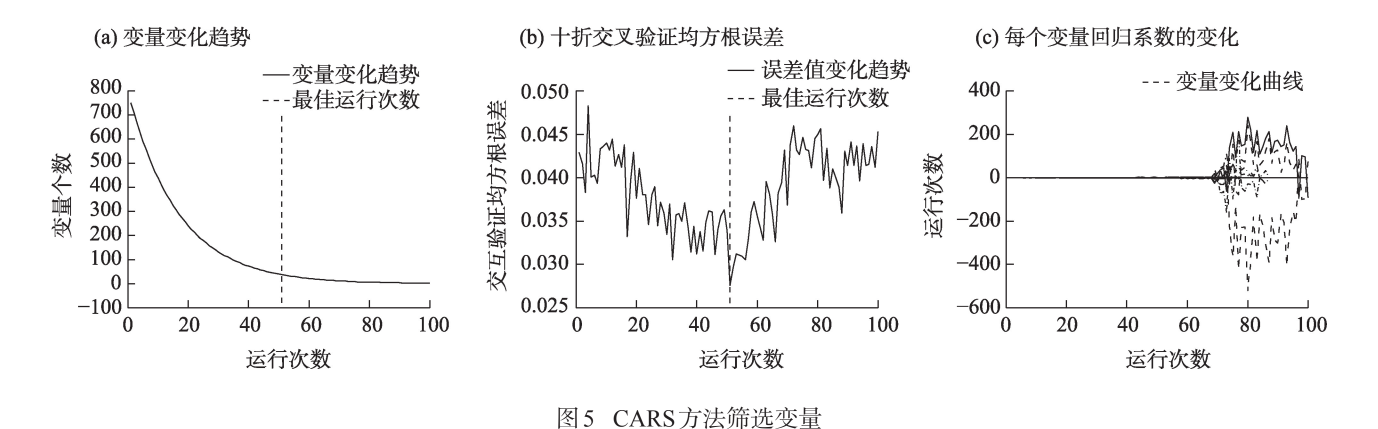
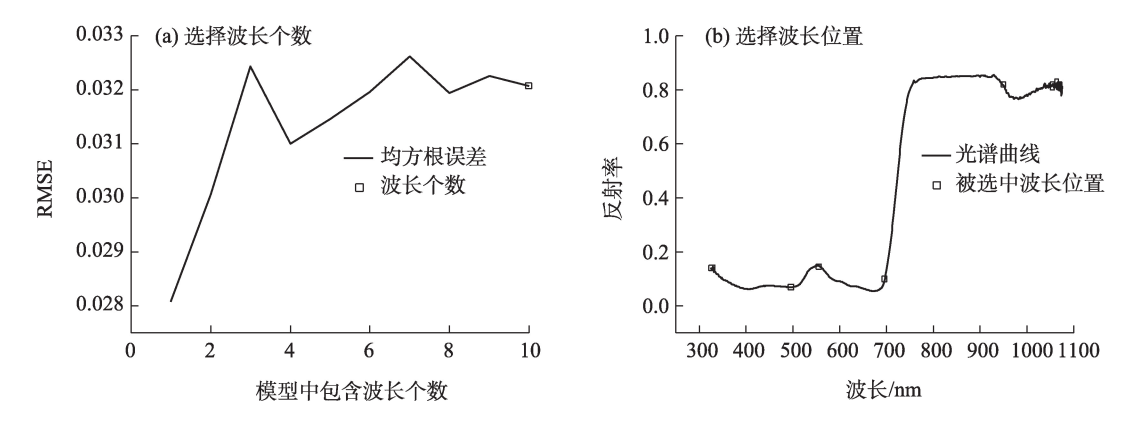


图5　CARS方法筛选变量

Fig. 5　Key variables selected by CARS method



注：RMSE为均方根误差。下同。

图6　SPA方法筛选变量

Fig. 6　Key variables selected by SPA method

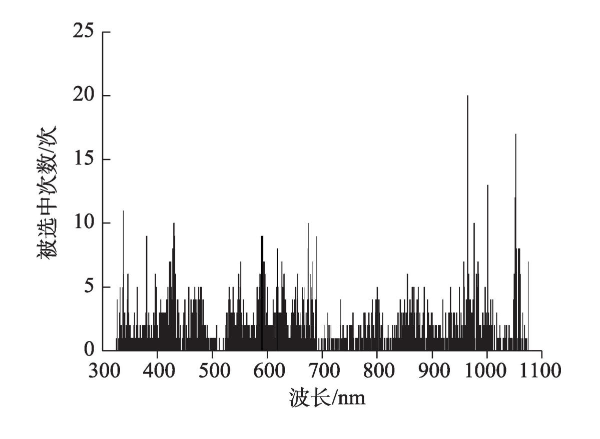


图7 GA方法筛选变量图

Fig. 7 Key variables selected by GA method

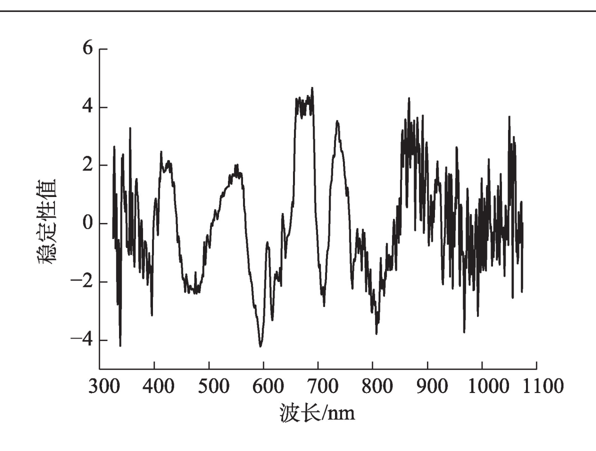
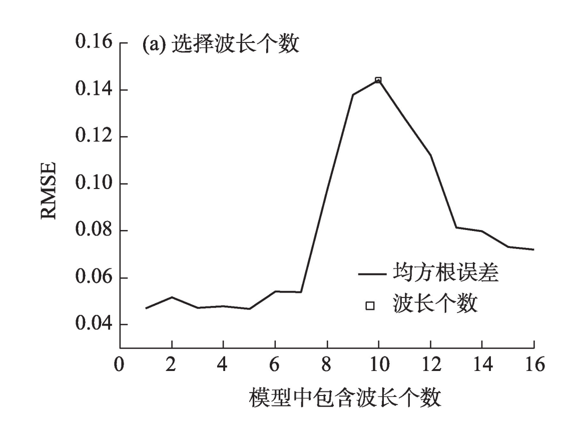


图8 MC-UVE方法筛选变量

Fig. 8 Key variables selected by MC-UVE method



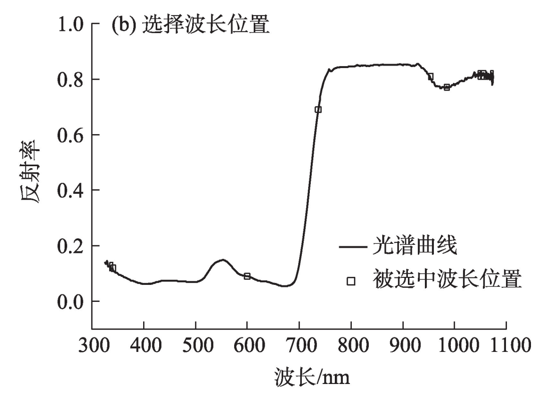


图9 CARS-SPA方法筛选变量

Fig. 9 Key variables selected by CARS-SPA method

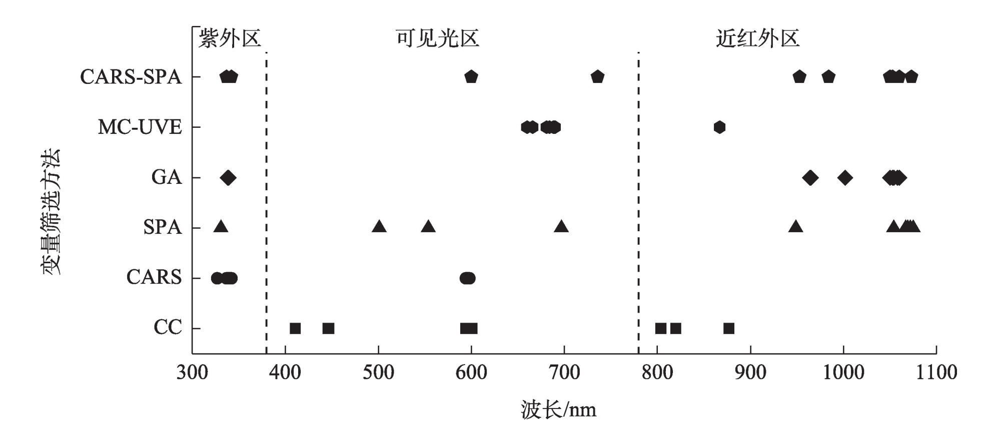


图10 不同变量筛选方法挑选特征波长分布

Fig. 10 Selection of characteristic wavelength distribution by different variable screening methods

并与SVR和RFR相比较。

对于CC分析法，经过分数阶微分处理后，通过0.01显著性水平检验且波段数量大于10个的阶次建立反演模型。如表3所示，选择R^2最大且RMSE最小的阶次进行建模分析，相比其他阶次的微分变换，1.3阶微分所构建模型的$R^2 \geq 0.881$，RMSE≤0.019，说明此阶次的微分处理效果较好，故选用1.3阶变换进行后续分析。

由表4可知，全波段中模型预测效果均较差，而SPA-WOA-RFR与CARS-SPA-WOA-RFR模型的建

表3 不同阶微分下对棉花叶片含水量的建模结果

Tab. 3 Modeling results of leaf water content of cotton under different order differential

微分阶数	算法	表达式	R^2	RMSE
0.7	WOA-RFR	y=0.6934x+0.2403	0.894	0.017
	SVR	y=0.4627x+0.4201	0.553	0.029
	RFR	y=0.5463x+0.3546	0.885	0.021
0.9	WOA-RFR	y=0.7086x+0.2290	0.846	0.018
	SVR	y=0.5242x+0.3706	0.650	0.026
	RFR	y=0.5545x+0.3491	0.882	0.021
1.1	WOA-RFR	y=0.7086x+0.2254	0.877	0.017
	SVR	y=0.6291x+0.2896	0.767	0.022
	RFR	y=0.6121x+0.3045	0.897	0.019
1.3	WOA-RFR	y=0.6778x+0.2522	0.927	0.016
	SVR	y=0.7505x+0.1953	0.881	0.016
	RFR	y=0.6206x+0.2976	0.898	0.019
1.7	WOA-RFR	y=0.6381x+0.2816	0.844	0.020
	SVR	y=0.7293x+0.2117	0.889	0.016
	RFR	y=0.5831x+0.3271	0.893	0.020
1.9	WOA-RFR	y=0.7061x+0.2293	0.851	0.018
	SVR	y=0.6169x+0.3022	0.732	0.023
	RFR	y=0.5676x+0.3392	0.880	0.021

注：R^2为决定系数；RMSE为均方根误差；WOA为鲸鱼优化算法；SVR为支持向量机回归；RFR为随机森林回归。下同。

表4 叶片含水量预测模型在建模集与验证集的R^2与RMSE

Tab. 4 R^2 and RMSE of leaf water content prediction models in calibration and validation sets

变量筛选方法	算法	建模集		验证集	
		R^2	RMSE	R^2	RMSE
全波段	WOA-RFR	0.573	0.029	0.480	0.030
	SVR	0.375	0.032	0.035	0.037
	RFR	0.583	0.030	0.021	0.032
CC	WOA-RFR	0.927	0.016	0.946	0.017
	SVR	0.881	0.016	0.908	0.016
	RFR	0.898	0.019	0.950	0.022
CARS	WOA-RFR	0.912	0.017	0.929	0.015
	SVR	0.688	0.040	0.977	0.033
	RFR	0.889	0.025	0.916	0.023
SPA	WOA-RFR	0.937	0.016	0.941	0.015
	SVR	0.398	0.034	0.757	0.024
	RFR	0.913	0.022	0.964	0.022
GA	WOA-RFR	0.622	0.028	0.647	0.025
	SVR	0.712	0.024	0.723	0.002
	RFR	0.889	0.024	0.858	0.001
MC-UVE	WOA-RFR	0.847	0.020	0.868	0.016
	SVR	0.882	0.015	0.883	0.015
	RFR	0.852	0.025	0.821	0.024
CARS-SPA	WOA-RFR	0.935	0.017	0.942	0.019
	SVR	0.326	0.036	0.568	0.029
	RFR	0.911	0.023	0.878	0.024

模集与验证集的 R^2 和RMSE差异较小，表明基于WOA优化的RFR模型较稳定。综合考虑数据集的 R^2 和RMSE，选择CC-WOA-RFR、CARS-WOA-RFR、SPA-WOA-RFR、GA-RFR、MC-UVE-SVR和CARS-SPA-WOA-RFR作为验证数据集Ⅱ的模型，如图11所示，所选择模型的拟合值均匀分布在1:1拟合线两侧，模型拟合效果较好。

为了验证模型稳定性，使用数据集Ⅱ对上述6个模型进行再次验证。模型结果如图12所示，CC-WOA-RFR模型结果与建模时相比 R^2 较低，表明该模型稳定性较差，不能有效估测棉花叶片含水量。而CARS-WOA-RFR、GA-RFR与MC-UVE-SVR模型

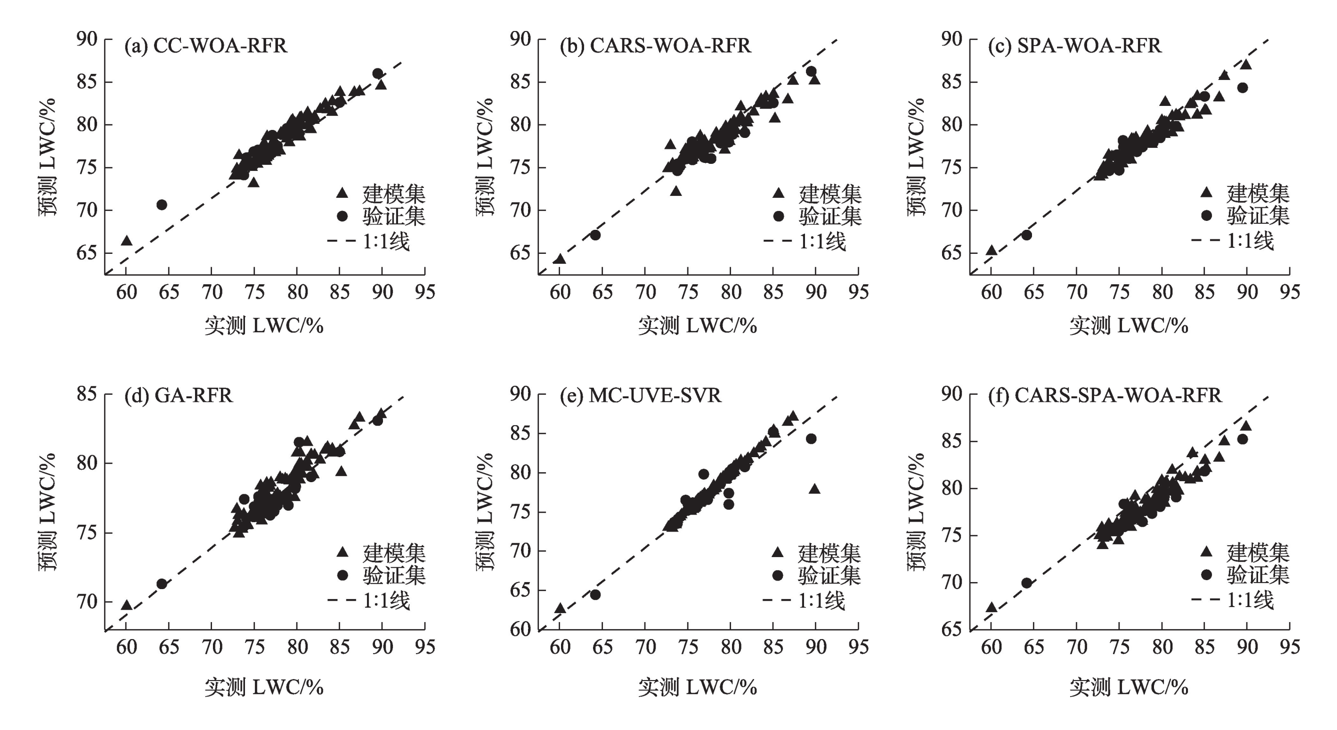


图11 数据集Ⅰ棉花叶片水分含量实测值与预测值散点图

Fig. 11 Data set Ⅰ scatter plot of measured and predicted cotton leaf water content

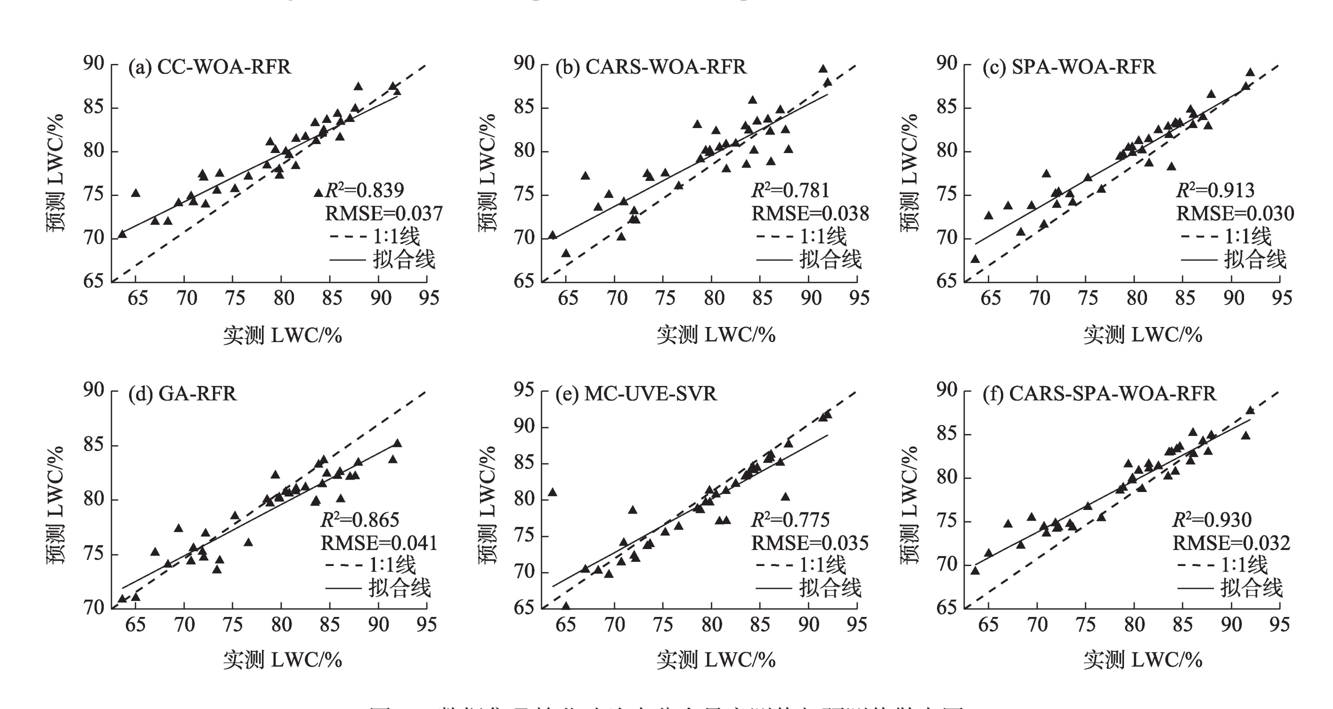


图12 数据集Ⅱ棉花叶片水分含量实测值与预测值散点图

Fig. 12 Data set II scatter plot of measured and predicted cotton leaf water content

R^2均较低，实测值与预测值误差较大。SPA-WOA-RFR与CARS-SPA-WOA-RFR模型反演精度较高，$R^2 \geqslant 0.913$，RMSE≤0.032。上述模型中，CARS-SPA-WOA-RFR模型R^2最高，可以较为准确地估测棉花叶片水分含量。

3 讨论

光谱预处理是改善数据质量和提升建模精度的必要手段，本文使用分数阶微分对棉花叶片高光谱数据进行处理，显著提高了光谱数据与叶片水分含量的相关性，这与于雷等[12]和吾木提·艾山江等[29]对土壤和小麦叶片光谱处理后相关性水平得到提升的研究结果一致。

高光谱数据由于波段较多，包含了大量与水分不相关且低贡献度的波段，全波段建模中效果较差。因此，本文通过6种波长变量筛选方法的计算，有效筛选了特征波长，降低了数据冗余，建模效果得到提升，这与Zhang等[30]、Han等[31]与Jia等[32]使用CARS、SPA等方法选择特征波长中所得结论一致。特别是本文通过CARS与SPA方法的耦合更加有效地筛选了特征波长，最终所构建模型效果最优，这与于雷等[12]得到的结论相一致。

不同的机器学习模型对于同一数据的预测效果会有部分差异[31]，基于WOA-RFR的棉花叶片水分含量反演模型精度优于SVR和RFR模型，且通过独立样本集检验后模型仍较稳定，能取得较好的反演效果。与Li等[14]、Zhou等[15]和Mohammadi等[33]的研究结果相同，即基于WOA算法改进后的模型预测精度均能得到有效提高。

本研究使用分数阶微分对原始光谱数据进行预处理，有效降低了环境对光谱数据的影响，但是野外数据采集依然受到土壤、大气和周边冠层等影响，以及存在地域差异，不同地区的棉花叶片高光谱特征存在略微差别。导致所选择的特征波段与前人研究中水分的敏感波段有部分差异，比如CC、CARS和MC-UVE方法所选择的波段位置出现一定的偏移，多集中于紫外区和可见光区。此外，本研究使用2个不同生育期的数据集对模型进行反复验证，一定程度上克服了以往研究中数据集单一的缺陷。但最终所估测结果出现不同程度的"低值高估与高值低估"现象，在未来的研究中需要借助更加优化的机器学习算法对本模型进行验证与校正，进一步提高模型的稳定性与适用性。

4 结论

（1）分数阶微分的光谱预处理方法可以提高相关性水平。其中，0.7阶与0.9阶处理效果比较明显。

（2）不同的特征波段筛选方法所得波段数量与位置均有差异。其中，MC-UVE所得变量最少（8个），CARS所得最多（38个）。SPA、GA、CARS-SPA所筛选波段位置较为一致，CC与MC-UVE差异较大。

（3）WOA-RFR模型在反演中取得了较好的效果。通过数据集Ⅰ和Ⅱ的验证，CARS-SPA-WOA-RFR模型反演精度较高。模型预测值R^2=0.93，RMSE=0.032，表明该模型针对不同生长期和不同品种棉花叶片含水量的预测均可以取得较好的精度。

参考文献（References）

[1] 刘文静, 范永胜, 董彦琪, 等. 我国棉花生产现状分析及建议[J]. 中国种业, 2022(1): 21–25. [Liu Wenjing, Fan Yongsheng, Dong Yan Qi, et al. Analysis and suggestions on the current situation of cotton production in China[J]. China Seed Industry, 2022(1): 21–25.]

[2] 马春玥, 买买提·沙吾提, 依尔夏提·阿不来提, 等. 新疆棉花种植业地理集聚特征及影响因素研究[J]. 作物学报, 2019, 45(12): 1859–1867. [Ma Chunyue, Sawut Mamat, Ablet Ershat, et al. Characteristics and influencing factors of geographical agglomeration of cotton plantation in Xinjiang[J]. Acta Agronomica Sinica, 2019, 45(12): 1859–1867.]

[3] 孙俊, 丛孙丽, 毛罕平, 等. 基于高光谱的油麦菜叶片水分CARS-ABC-SVR预测模型[J]. 农业工程学报, 2017, 33(5): 178–184. [Sun Jun, Cong Sunli, Mao Hanping, et al. CARS-ABC-SVR model for predicting leaf moisture of leaf-used lettuce based on hyperspectral[J]. Transactions of the Chinese Society of Agricultural Engineering, 2017, 33(5): 178–184.]

[4] Junttila S, Hölttä T, Saarinen N, et al. Close-range hyperspectral spectroscopy reveals leaf water content dynamics[J]. Remote Sensing of Environment, 2022, 277: 113071, doi: 10.20944/preprints202108.0497.v1.

[5] Li L, Ustin S L, Riano D. Retrieval of fresh leaf fuel moisture content using genetic algorithm partial least squares (GA-PLS) modeling[J]. IEEE Geoscience and Remote Sensing Letters, 2007, 4(2): 216–220.

[6] Sun J, Zhou X, Hu Y G, et al. Visualizing distribution of moisture content in tea leaves using optimization algorithms and NIR hyperspectral imaging[J]. Computers and Electronics in Agriculture, 2019, 160: 153–159.

[7] Li X L, Wei Z X, Peng F F, et al. Estimating the distribution of

chlorophyll content in CYVCV-infected lemon leaf using hyperspectral imaging[J]. Computers and Electronics in Agriculture, 2022, 198: 107036, doi: 10.1016/j.compag.2022.107036.

[8] 杨宝华, 陈建林, 陈林海, 等. 基于敏感波段的小麦冠层氮含量估测模型[J]. 农业工程学报, 2015, 31(22): 176-182. [Yang Baohua, Chen Jianlin, Chen Linhai, et al. Estimation model of wheat canopy nitrogen content based on sensitive bands[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(22): 176-182.]

[9] 张文旭, 佟炫梦, 周天航, 等. 基于高光谱成像的棉花叶片氮素含量遥感估测[J]. 沈阳农业大学学报, 2021, 52(5): 586-596. [Zhang Wenxu, Tong Xuanmeng, Zhou Tianhang, et al. Remote sensing estimation of cotton leaf nitrogen content based on hyperspectral imaging[J]. Journal of Shenyang Agricultural University, 2021, 52(5): 586-596.]

[10] 易翔, 张立福, 吕新, 等. 基于无人机高光谱融合连续投影算法估算棉花地上部生物量[J]. 棉花学报, 2021, 33(3): 224-234. [Yi Xiang, Zhang Lifu, Lü Xin, et al. Estimation of cotton aboveground biomass based on unmanned aerial vehicle hyperspectral and successive projections algorithm[J]. Cotton Science, 2021, 33(3): 224-234.]

[11] 陈鹏. 基于无人机多源遥感的马铃薯叶绿素含量反演机理及模型构建[D]. 焦作: 河南理工大学, 2019. [Chen Peng. Retrieval mechanism and model construction of chlorophyll content in potato based on multi-source remote sensing of unmanned aerial vehicle[D]. Jiaozuo: Henan Polytechnic University, 2019.]

[12] 于雷, 洪永胜, 周勇, 等. 高光谱估算土壤有机质含量的波长变量筛选方法[J]. 农业工程学报, 2016, 32(13): 95-102. [Yu Lei, Hong Yongsheng, Zhou Yong, et al. Wavelength variable selection methods for estimation of soil organic matter content using hyperspectral technique[J]. Transactions of the Chinese Society of Agricultural Engineering, 2016, 32(13): 95-102.]

[13] 王佳文, 彭杰, 纪文君, 等. 基于电磁感应数据的南疆棉田土壤pH反演研究[J]. 干旱区研究, 2022, 39(4): 1293-1302. [Wang Jiawen, Peng Jie, Ji Wenjun, et al. Soil pH inversion based on electromagnetic induction data in cotton field of southern Xinjiang[J]. Arid Zone Research, 2022, 39(4): 1293-1302.]

[14] Li L L, Sun J, Tseng M L, et al. Extreme learning machine optimized by whale optimization algorithm using insulated gate bipolar transistor module aging degree evaluation[J]. Expert Systems with Applications, 2019, 127: 58-67.

[15] Zhou J, Zhu S Li, Qiu Y G, et al. Predicting tunnel squeezing using support vector machine optimized by whale optimization algorithm[J]. Acta Geotechnica, 2022, 17: 1343-1366.

[16] Zhao F, Li W D. A combined model based on feature selection and WOA for $PM_{2.5}$ concentration forecasting[J]. Atmosphere, 2019, 10(4): 223, doi: 10.3390/atmos10040223.

[17] 苏毅, 王克如, 李少昆, 等. 棉花植株水分含量的高光谱监测模型研究[J]. 棉花学报, 2010, 22(6): 554-560. [Su Yi, Wang Keru, Li Shaokun, et al. Monitoring models of the plant water content based on cotton canopy hyperspectral reflectance[J]. Cotton Science, 2010, 22(6): 554-560.]

[18] 王强, 易秋香, 包安明, 等. 棉花冠层水分含量估算的高光谱指数研究[J]. 光谱学与光谱分析, 2013, 33(2): 507-512. [Wang Qiang, Yi Qiuxiang, Bao Anming, et al. Discussion on hyperspectral index for the estimation of cotton canopy water content[J]. Spectroscopy and Spectral Analysis, 2013, 33(2): 507-512.]

[19] 赵巧珍, 丁建丽, 韩礼敬, 等. MODIS和Landsat时空融合影像在土壤盐渍化监测中的适用性研究——以渭干河—库车河三角洲绿洲为例[J]. 干旱区地理, 2022, 45(4): 1155-1164. [Zhao Qiaozhen, Ding Jianli, Han Lijing, et al. Exploring the application of MODIS and Landsat spatiotemporal fusion images in soil salinization: A case of Ugan River-Kuqa River Delta Oasis[J]. Arid Land Geography, 2022, 45(4): 1155-1164.]

[20] 玉苏甫·买买提, 吐尔逊·艾山, 买合皮热提·吾拉木. 新疆渭-库绿洲棉花种植面积遥感监测研究[J]. 农业现代化研究, 2014, 35(2): 240-243. [Mamat Yusup, Hasan Tursun, Gulam Magpirat. Remote sensing of cotton plantation areas monitoring in delta oasis of Ugan-Kucha River, Xinjiang[J]. Research of Agricultural Modernization, 2014, 35(2): 240-243.]

[21] 刘帆. 分数阶微分算法在医学超声弹性图像去噪中的应用研究[D]. 昆明: 昆明理工大学, 2018. [Liu Fan. Application of fractional differential algorithm in medical ultrasonic elastic image denoising[D]. Kunming: Kunming University of Science and Technology, 2018.]

[22] 李长春, 施锦锦, 马春艳, 等. 基于小波变换和分数阶微分的冬小麦叶绿素含量估算[J]. 农业机械学报, 2021, 52(8): 172-182. [Li Changchun, Shi Jinjin, Ma Chunyan, et al. Estimation of chlorophyll content in winter wheat based on wavelet transform and fractional differential[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(8): 172-182.]

[23] Li H D, Liang Y Z, Xu Q S, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. Analytica Chimica Acta, 2009, 648(1): 77-84.

[24] Zhang J K, Rivard B, Rogge D M. The successive projection algorithm (SPA), an algorithm with a spatial constraint for the automatic search of endmembers in hyperspectral data[J]. Sensors, 2008, 8(2): 1321-1342.

[25] 程介虹, 陈争光, 衣淑娟. 最小相关系数的多元校正波长选择算法[J]. 光谱学与光谱分析, 2022, 42(3): 719-725. [Cheng Jiehong, Chen Zhengguang, Yi Shujuan. Wavelength selection algorithm based on minimum correlation coefficient for multivariate calibration[J]. Spectroscopy and Spectral Analysis, 2022, 42(3): 719-725.]

[26] 宋相中. 近红外光谱定量分析中三种新型波长选择方法研究[D]. 北京: 中国农业大学, 2017. [Song Xiangzhong. Research of three new wavelength selection methods in near infrared spectroscopy quantitative analysis area[D]. Beijing: China Agricultural

University, 2017.]

[27] Mirjalili S, Lewis A. The whale optimization algorithm[J]. Advances in Engineering Software, 2016, 95: 51–67.

[28] 李畸勇, 张伟斌, 赵新哲, 等. 改进鲸鱼算法优化支持向量回归的光伏最大功率点跟踪[J]. 电工技术学报, 2021, 36(9): 1771–1781. [Li Qiyong, Zhang Weibin, Zhao Xinzhe, et al. Global maximum power point tracking for PV array based on support vector regression optimized by improved whale algorithm[J]. Transactions of China Electrotechnical Society, 2021, 36(9): 1771–1781.]

[29] 吾木提·艾山江, 买买提·沙吾提, 马春玥. 基于分数阶微分和连续投影算法-反向传播神经网络的小麦叶片含水量高光谱估算[J]. 激光与光电子学进展, 2019, 56(15): 251–259. [Hasan Umut, Sawut Mamat, Ma Chunyue. Hyperspectral estimation of wheat leaf water content using fractional differentials and successive projection algorithm-back propagation neural network[J]. Laser & Optoelectronics Progress, 2019, 56(15): 251–259.]

[30] Zhang M J, Zhang S Z, Iqbal J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares[J]. Chemometrics and Intelligent Laboratory Systems, 2013, 128: 17–24.

[31] Han Z Z, Deng L M. Application driven key wavelengths mining method for aflatoxin detection using hyperspectral data[J]. Computers and Electronics in Agriculture, 2018, 153: 248–255.

[32] Jia M, Li W, Wang K K, et al. A newly developed method to extract the optimal hyperspectral feature for monitoring leaf biomass in wheat[J]. Computers and Electronics in Agriculture, 2019, 165: 104942, doi: 10.1016/j.compag.2019.104942.

[33] Mohammadi B, Mehdizadeh S. Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm[J]. Agricultural Water Management, 2020, 237: 106145, doi: 10.1016/j.agwat.2020.106145.

Estimation of leaf water content in upland cotton based on feature band selection and machine learning

CUI Jintao[1],　Mamat SAWUT[1,2,3]

(1. College of Geography and Remote Sensing Sciences, Xinjiang University, Urumqi 830046, Xinjiang, China; 2. Xinjiang Key Laboratory of Oasis Ecology, Xinjiang University, Urumqi 830046, Xinjiang, China; 3. Key Laboratory of Smart City and Environment Modelling of Higher Education Institute, Xinjiang University, Urumqi 830046, Xinjiang, China)

Abstract: It is critical to ensure timely and accurate monitoring of leaf water content (LWC) when assessing the growth status of cotton. To accurately estimate cotton LWC, hyperspectral data, and leaf water data from cotton leaves in the oasis of the Ugan River-Kuqa River Delta, Xinjiang, China, were selected and processed using fractional differentiation of raw spectra. The sample were analyzed through correlation coefficient analysis, competitive adaptive reweighted sampling (CARS), successive projections algorithm (SPA), genetic algorithm (GA), Monte Carlo uninformative variables elimination (MC-UVE), and a combination of CARS and SPA to filter the feature bands. The modeling of the LWC inversion was executed through random forest regression (RFR) based on the whale optimization algorithm (WOA), and independent samples were used for validation analysis. The results show that: (1) The disparities in the number and positions of the feature bands obtained using the different feature band screening methods are different, where the number of feature bands obtained through MC-UVE is 8 while CARS produced 38. The positions of the characteristic bands identified through the SPA, GA, and CARS-SPA methods are considerably consistent and fundamentally concentrated in the near-infrared range of 950–1050 nm. (2) The CARS-SPA-WOA-RFR model has the best inversion with an R^2 of 0.93 and a root mean square error of 0.032. This model can provide a decision basis for accurate and rapid monitoring of cotton drought and precision irrigation.

Key words: spectral; leaf water content; feature band selection; machine learning